Literature Survey

106121018 106121086 106121116

1. Web-Page Content Classification on Entropy Classifiers using Machine Learning

https://ieeexplore.ieee.org/document/10080462

In growing demand of technology and its authentication detection manages a key role via web servers and browser tools. Various classification techniques are proposed to manage systematic queries of web pages. Discusses a primary classification and feature extraction of web pages reviews. The author insights a detailed description towards classification of content and file types based on first layer attributes. The paper provides details about intermediate and third party platforms for server management and classification. The features play an important role for interconnected web pages such as keywords or key sentences, thus demonstrates systematic classification of features based on extraction. In various algorithms and supporting techniques such as SVM, KNN are elaborately discussed. These methodologies are typical and have seen various downfalls with increased scenario of web page count. An approach of web mining and classification was based on Linked Information Categories (LIC) providing a clear objective based classification of web pages. These patterns are higher in performance ratio and has shown a new benchmark for web classification.

A machine learning-based auto classification of web pages is discussed by a team of Osaka University, the paper focuses on error rate with respect to classification rate. Henceforth a detailed terminology of content-based classification is projected. From all stated papers, a noted scenario of linking information is underperformed and hence the proposed methodology in section 3 demonstrates a hierarchical-based classification.

Improvements:
> Enhance text preprocessing by removing HTML tags, punctuation, and special characters, and consider stemming or lemmatization to reduce word variations.

>Experiment with more advanced text representation techniques, such as TF-IDF, Word Embeddings (Word2Vec, GloVe), or pre-trained language models like BERT, to capture semantic meaning and context of words more effectively.

>Incorporate domain-specific features or metadata if available, as they can provide valuable information for classification.

>If webpages can belong to multiple categories simultaneously, adapt the classification approach to handle multilabel classification problems.

2. SafeBrowse: A new tool for strengthening and monitoring the security configuration of web browsers

https://ieeexplore.ieee.org/document/7479263

The growing complexity of direct cyberattacks on users' computers, coupled with the increasing number of antivirus programs, operating systems, and various defense mechanisms, has led cybercriminals to concentrate their efforts on exploiting web browsers and related technologies. Their target audience includes the vast majority of internet users due to the widespread use of web browsers on machines connected to the internet. This paper discusses the factors that make web browsers an attractive target for cybercriminals and emphasizes the need for enhanced browser security, especially with the rise of online activities like email, e-commerce, and e-learning.

The attackers' primary goal is to execute arbitrary code on a remote user's operating system, gaining control of their machine and accessing the stored information. Various methods are employed to compromise web browsers, including exploiting source code vulnerabilities, injecting malicious code into web pages, URL spoofing, and other techniques that destabilize browsers to facilitate control.

For internet users, web browsers serve as gateways to their privacy, and the increasing reliance on online transactions necessitates bolstering browser security. Despite efforts by browser designers and developers to enhance security by implementing numerous parameters, users' lack of awareness remains a major vulnerability that hackers exploit.

This paper introduces a new tool designed to enhance browser security by mitigating misconfiguration risks. The tool focuses on monitoring and managing the configurations of popular browsers installed on users' machines. The paper is structured as follows: Section II discusses various types of web browser attacks, Section III reviews prior work related to browser security, Section IV details the methodology and approach for monitoring and securing browser configurations, and Section V presents a discussion of the proposed solution.

Improvements:

>The algorithm relies on manual edits to configuration files or the Windows registry, which can be error-prone and may not be suitable for all users.Develop a mechanism to automatically configure security settings based on best practices or user preferences, rather than relying on manual edits to configuration files or registries.
> The algorithm is limited to Windows operating systems, which restricts its applicability to other platforms like macOS and Linux.

>Consider implementing real-time monitoring of browsing history and downloads to detect suspicious activity as it happens, providing better security.

3. Website categorization via design attribute learning

https://www.sciencedirect.com/science/article/pii/S016740482100136X

Introduces the context of cybersecurity with a focus on human factors, particularly human vulnerability to cyberattacks, and the challenges posed by crack websites distributing malware. The survey highlights the importance of identifying and categorizing websites based on their design features to improve cybersecurity mechanisms. Below is a summary of the key points and contributions of the literature:

Challenges of Crack Websites:

Crack websites distribute malware, evading antipiracy measures. Over 50% of these sites contain malicious software, posing a severe cybersecurity challenge.

Role of Visual Design in Identifying Crack Websites:

Visual design elements can serve as signals for crack websites, striking a balance between evading detection and attracting visitors. This study suggests that design features may contain consistent visual characteristics unique to crack websites.

Proposed Website Assessment Scheme:

The scheme focuses on extracting and analyzing website design features, including website address, element tags, area, keywords, text length, and text font rate. It aims to create a dataset for learning and identification using machine learning algorithms.

Experimental Settings and Evaluation:

Two experiments validate the proposed method. Experiment 1 identifies crack websites with an average accuracy of 90.7% using a decision tree classifier based on design features. Experiment 2 scales up to a larger dataset, including malicious websites, and uses various machine learning models for classification.

Results of Experiment 1:

The J48 decision tree classifier based on design features achieves an average accuracy of 90.7% in classifying crack websites, demonstrating the effectiveness of design features for classification.

Improvements:
>The proposed application receives a list of URLs as a flat input file.We can use live data, browsing history etc...
>Consider using more advanced machine learning models, such as gradient boosting (e.g., XGBoost or LightGBM) or deep learning models (e.g., convolutional neural networks for webpage image analysis), to potentially improve classification accuracy.

4. Web page classification based on heterogeneous features and a combination of multiple classifiers

https://link.springer.com/article/10.1631/FITEE.1900240

In this study, a novel web page classification method is proposed, aiming for precise classification by leveraging both textual and structural features of web pages. Unlike traditional methods that focus solely on textual features, this approach exploits the tree-like structure of HTML tags to capture the structural characteristics of web pages. The method combines heterogeneous textual and structural features by converting them into vectors and then fusing them.

To evaluate the reliability of classification results, the concept of "confidence" is introduced. Confidence is used to measure the trustworthiness of classification outcomes based on the predicted scores of different classifiers. The higher the confidence, the more reliable the classification result is considered.

Multiple classifiers, including LSTM (Long Short-Term Memory) networks and SVM (Support Vector Machines), are employed in this method. These classifiers each offer distinct advantages, and their outputs are combined using confidence-based decision strategies such as voting and direct output.

Experimental results demonstrate the effectiveness of this approach on various datasets, including the Amazon dataset, 7-web-genres dataset, and DMOZ dataset. The proposed method achieves higher classification accuracies compared to existing web page classification algorithms that rely solely on textual features. Additionally, the fusion of structural features with textual features contributes to the improved accuracy.

In summary, this study presents a comprehensive approach to web page classification, harnessing both textual and structural features and combining multiple classifiers based on confidence. This approach outperforms existing methods and offers a promising avenue for enhancing web page classification accuracy.

Improvements:

>Regular Updates:
Stay up to date with changes in web page structures, technologies, and trends to ensure the model remains effective over time.

>User Feedback Loop:
Implement a feedback loop that allows users to provide feedback on misclassified web pages. This feedback can be used to continually fine-tune and improve the classification model.

5. A genetic algorithm based focused Web crawler for automatic webpage classification

https://ieeexplore.ieee.org/document/8372223

Explores the application of Genetic Algorithms (GA) in the context of webpage classification for web crawling and search engines. Here is a summary of the key points:

- Web Crawling and Search Engines: The survey introduces search engines as programs designed to extract information from the web through web crawling, indexing, and searching. Web crawlers are highlighted as essential for systematically collecting web pages from the World Wide Web.
- Focused Web Crawlers: Focused web crawlers are discussed as specialized tools designed to collect web pages relevant to specific topics or categories. They consist of key components: a classifier, a distiller, and a crawler with priority control.
- Web Page Classification: The survey emphasizes the importance of web page classification within focused web crawling. It is described as a supervised learning problem, and various machine learning algorithms, including GA, are mentioned as approaches for classification.
- Related Work: Prior research on the use of GA in web page classification is summarized. This includes optimizing document descriptors, feature selection, query term weight assignment, and rule-based classification techniques.
- Web Page Classification Approaches: Different approaches to web page classification are outlined, including on-page, neighbor-based, URL-based, and algorithm-based methods.
- Proposed Genetic Algorithm: The survey introduces a GA-based system for web page classification, consisting of feature extraction, a GA-based classifier, and a classification process. It explains the initialization of GA parameters and the termination condition based on convergence.
- Classification Process: In the classification phase, web pages are crawled and filtered, and each page is represented as a document with features. Cosine similarity is used to categorize web pages as relevant or not.
- Experimental Results: The survey details experiments conducted on datasets from Wikipedia and the regular web. Performance metrics such as precision, recall, accuracy, and F1 score are used to evaluate the proposed system's performance, which is found to outperform a baseline GA-based focused crawler, especially on the regular web dataset.

Overall, the literature survey explores the use of GA as a powerful tool for web page classification in the context of focused web crawling, providing insights into its application and its performance compared to existing approaches.

Improvement:

>Investigate alternative initialization strategies for chromosome weights. Instead

of random initialization between 0 and 1, consider heuristics or techniques like K-means clustering to provide a better starting point for the optimization process

>Instead of using a fixed threshold, consider using dynamic thresholding techniques based on the distribution of fitness scores in each generation. Adaptive thresholds can adapt to changing data characteristics.

6.**Web Page Classification Algorithm Based on Deep Learning**

https://www.hindawi.com/journals/cin/2022/9534918/

In the era of 5G technology and the explosion of online data, effective web page classification is essential for content understanding and recommendation systems. Deep Learning (DL) has significantly advanced image and speech processing, revolutionizing various industrial applications, including recommendation systems. This research focuses on Web Page Classification Algorithms (WPCA) based on DL, aiming to enhance their accuracy and efficiency.

The article introduces three key innovations. Firstly, it proposes a keyword weight calculation method to improve WPCA accuracy by reducing the influence of high-frequency words and adjusting the weights of low-frequency words. Secondly, it presents a Chinese web page classification method that calculates similarity between the text to be classified and predefined class templates, making classification rules-based decisions. Finally, to enhance DL's learning rate, the article explores adaptive parameter optimization algorithms that automatically adjust learning rates, improving WPCA efficiency.

Comparative experiments between DL-based WPCA and traditional algorithms show that DL-based WPCA is significantly faster and consumes less memory. The DL-based WPCA takes only 354 s in terms of time expenditure compared to 2436 s for the traditional algorithm. Memory overhead is reduced to 6.35 s compared to 186.25 s for the traditional method.

The research's practical implications are substantial, benefiting directory website maintenance, topic browser construction, and enhancing user experiences in search engines and recommendation systems. This study demonstrates the potential of DL-based WPCA in handling large-scale web page classification tasks efficiently and accurately, contributing to the management and organisation of ever-increasing online data.

Improvement:

Hybrid Models: Consider hybrid models that combine DL-based approaches with traditional machine learning algorithms. Hybrid models can leverage the strengths of both paradigms, potentially leading to improved classification performance.

User Feedback Integration: If possible, incorporate user feedback into the model's training and evaluation process. User interactions and feedback can help continuously improve the recommendation and classification accuracy.